Empirical Article

Advances in Methods and Practices in Psychological Science
October-December 2023, Vol. 6, No. 4, pp. 1–17

DOI: 10.1177/25152459231205832
www.psychologicalscience.org/AMPPS

Open-Science Guidance for Qualitative Research: An Empirically Validated Approach for De-Identifying Sensitive Narrative Data

Rebecca Campbell[1], McKenzie Javorka[2], Jasmine Engleton[1], Kathryn Fishwick[3], Katie Gregory[1], and Rachael Goodman-Williams[3]

[1]Department of Psychology, Michigan State University, East Lansing, Michigan, [2]The Rural Institute for Inclusive Communities, University of Montana, Missoula, Montana, and [3]Department of Psychology, Wichita State University, Wichita, Kansas

Abstract

The open-science movement seeks to make research more transparent and accessible. To that end, researchers are increasingly expected to share de-identified data with other scholars for review, reanalysis, and reuse. In psychology, open-science practices have been explored primarily within the context of quantitative data, but demands to share qualitative data are becoming more prevalent. Narrative data are far more challenging to de-identify fully, and because qualitative methods are often used in studies with marginalized, minoritized, and/or traumatized populations, data sharing may pose substantial risks for participants if their information can be later reidentified. To date, there has been little guidance in the literature on how to de-identify qualitative data. To address this gap, we developed a methodological framework for remediating sensitive narrative data. This multiphase process is modeled on common qualitative-coding strategies. The first phase includes consultations with diverse stakeholders and sources to understand reidentifiability risks and data-sharing concerns. The second phase outlines an iterative process for recognizing potentially identifiable information and constructing individualized remediation strategies through group review and consensus. The third phase includes multiple strategies for assessing the validity of the de-identification analyses (i.e., whether the remediated transcripts adequately protect participants' privacy). We applied this framework to a set of 32 qualitative interviews with sexual-assault survivors. We provide case examples of how blurring and redaction techniques can be used to protect names, dates, locations, trauma histories, help-seeking experiences, and other information about dyadic interactions.

Keywords

open science, data sharing, data archiving, qualitative interview, de-identification, dyadic data

Received 4/27/23; Revision accepted 8/15/23

Over the past decade, the open-science movement has become a transdisciplinary effort to make the processes and outcomes of research more transparent and accessible to other scholars, policymakers, and the public writ large (Nosek et al., 2015; Vicente-Saez & Martinez-Fuentes, 2018). To that end, a recommended best practice is data sharing, whereby researchers make their data-collection instruments, data sets, and analyses available in public archives (Hesse, 2018; Meyer, 2018). Data sharing has many purposes, some of which focus on ensuring accuracy, others on promoting new discoveries. First, data sharing allows researchers to assess the reproducibility of a project: whether others can follow the same procedures with the same data to reproduce the same results (American Statistical Association, 2017).

Corresponding Author:
Rebecca Campbell, Department of Psychology, Michigan State University, East Lansing, Michigan
Email: rmc@msu.edu

Verifying reproducibility is critical for correcting or preventing errors in published works (McNutt, 2014). Second, data sharing helps researchers assess the replicability of a study's findings: whether others can follow the same procedures to collect new data with new participants and obtain the same results (Pashler & Wagenmakers, 2012). Within the discipline of psychology, replication projects are seen as critical for restoring trust in psychological science and promoting a culture of accountability (De Boeck & Jeon, 2018). Third, data sharing encourages the reuse of existing data to pursue novel questions (DuBois et al., 2018). Many funders now require researchers to archive their data for others to reuse to maximize the return on investment (DuBois et al., 2018; Mozersky, Parsons, et al., 2020).

To date, much of the discourse about data sharing has focused on quantitative data; however, proponents of open science contend that the movement's key aim to promote transparency is relevant across all modes of inquiry, and thus qualitative data should also be shared (see Field et al., 2021; Siegel et al., 2021 for reviews). Many qualitative scholars have cautioned against uncritical adoption of practices developed for a strikingly different type of research (Feldman & Shaw, 2019; Parry & Mauthner, 2004; Tsai et al., 2016). Qualitative inquiry has distinct ontological and epistemological assumptions about the existence of an objective reality, and its methodology is relational and intersubjective (Lincoln et al., 2018). The data to be shared are not numbers but, rather, words, images, and/or sounds. Given these essential differences, many qualitative scholars remain skeptical about whether reproducibility, replication, and/or reuse are sensible goals for this type of inquiry (Bennett, 2021; Brabeck, 2021; Feldman & Shaw, 2019; Parry & Mauthner, 2004). However, other qualitative researchers maintain that the open-science movement's guiding principles can be reimagined and tailored to narrative inquiry (Class et al., 2021; DuBois et al., 2018; Steltenpohl et al., 2023). As these debates have been unfolding in the literature, the expectations for greater transparency have grown stronger. Qualitative researchers are increasingly expected by academic journals, professional associations, and funding agencies to make their data available to others—even though there is not widespread agreement within the academic community on how to share narrative data safely and ethically.

In this article, we share our experiences of being caught between that proverbial rock and a hard place. We conducted a narrative-interview study with sexual-assault survivors about their interactions with the criminal legal system, which yielded deeply troubling stories of abuse—by the perpetrators and by the systems that were supposed to help these survivors but did not. This study was funded by a U.S. Department of Justice grant that required all grantees to submit de-identified data to the National Archive of Criminal Justice Data. We were quite concerned about the risks associated with sharing these data—and archiving was mandated by the funding agency. We acknowledge there are deep epistemological divides on this issue—and we had daunting ethical and methodological issues to resolve. We decided to approach this challenge as an opportunity to study the process of remediating highly sensitive narrative data. In this "meta study," we systematically tracked the procedures, decisions, and coding processes we employed to protect the identities of our research participants. In this article, we share these methods and case examples to illustrate the complexities of open-science practices in qualitative research and specific strategies for how these challenges can be addressed. We begin by reviewing key epistemological, ethical, and methodological concerns that researchers must address if they intend to share qualitative data. It is beyond the scope of this article to unpack each of these sizable dilemmas in depth (for reviews, see DuBois et al., 2018; Field et al., 2021; Steltenpohl et al., 2023), so in our review, we highlight the tensions between promoting transparency and protecting privacy.

Epistemological Tensions With Sharing Qualitative Data

The open-science movement generally and the practice of data sharing specifically are often described in conjunction with reproducibility and replication, and in some scientific circles, these practices are defined as essentialist components of high-quality research (Siegel et al., 2021). Such framing has drawn criticism from qualitative scholars because narrative inquiry is grounded in constructivist or critical epistemologies, without ontological assumptions of an objective reality (Lincoln et al., 2018). Qualitative research seeks to understand how people interpret their lived experiences, recognizing that such knowledge is bounded by history and context (Field et al., 2021; Steltenpohl et al., 2023). There is no expectation that a different analyst would necessarily interpret data the same way or that another researcher pursuing the same or similar research questions in another setting would obtain the same results (Tsai et al., 2016). Therefore, sharing data to meet a goal that is not actually a goal of this type of scholarship is unwarranted (Bennett, 2021; Brabeck, 2021; Feldman & Shaw, 2019).

However, open science is fundamentally about transparency—a call to scholars to be clearer about their processes so that diverse stakeholder audiences can understand their research (Field et al., 2021; Kapiszewski & Karcher, 2021). If quantitative scholars concretize

these ideals as reproducibility and replication, there is no reason qualitative scholars must follow suit (Karcher et al., 2021). As Kapiszewski and Karcher (2021) noted, "transparency is not an all-or-nothing proposition and can be pursued in many different ways" (p. 285). It is not epistemologically incongruent to challenge long-standing norms of methodological opacity and to prompt researchers for more detail about how their findings were generated and the contexts that bound that knowledge (Karcher et al., 2021; Rallis, 2015). Therefore, sharing data to promote transparency is an important endeavor in qualitative research.

Qualitative scholars have also questioned whether sharing narrative data to promote reuse and secondary analysis is truly feasible (Feldman & Shaw, 2019; Parry & Mauthner, 2004). Narrative data are cocreated between researchers and participants, and those unique interpersonal relationships shape what is disclosed and how that information is interpreted (Lincoln et al., 2018). Furthermore, when qualitative methods are used in participatory action research projects, researchers have long-standing relationships not only with participants but also with organizations and the community writ large (Kemmis et al., 2015). Prolonged engagement is a hallmark of qualitative research (Lincoln & Guba, 1985), and those months or years of interactions create the contextual foundation that influences all aspects of data collection, analysis, and interpretation. Secondary analysts do not have these in vivo experiences, so they may misunderstand or misinterpret narrative data, which limits the utility of qualitative data sets for new research questions (Feldman & Shaw, 2019; Parry & Mauthner, 2004).

However, the richness of narrative data begs the question of whether more could be learned and whether there are additional discoveries to be made through data sharing. In a recent review of secondary qualitative-data-analysis projects, Ruggiano and Perry (2019) found that most reuse studies were conducted by the original researchers, typically to pursue new questions that were unexplored in their prior analyses. Thus, narrative data do lend themselves to reuse—but possibly only for people who have a deep understanding of the data and their history. On the other hand, perhaps the research context is not so complicated and unknowable. Mozersky et al.'s (2022) content analysis of 100 qualitative health studies found that the data were typically collected in a single interaction, usually less than 1 hr in length, which suggests that not all qualitative studies have a deep relational context that secondary analysts could not understand. Furthermore, qualitative researchers can share their lived experiences of collecting their data. Kapiszewski and Karcher (2021) recommended that qualitative researchers publish/post supplemental methodological appendices that tell these stories and provide key contextual details secondary users may need to understand the data more fully.

Ethical Issues to Address When Sharing Qualitative Data

Whether the interaction between a researcher and a participant is long or short, qualitative research is a relational experience, and it stands to reason that some people may not want what they felt comfortable disclosing to a specific researcher shared with others. Thus, data sharing raises ethical questions about participants' agency and control of their information. Formative studies that have explored how research participants feel about data sharing have found they are generally agreeable because they want to maximize what can be learned from their experiences to help others (Campbell, Goodman-Williams, Engleton, et al., 2023; Kuula, 2011; Mozersky, Parsons, et al., 2020; VandeVusse et al., 2022; Yardley et al., 2014). However, participants emphasized that researchers must seek informed consent for data sharing; consent to participate in the study does not give researchers implicit permission to share the data collected in that study (for consent-language options, see Kaiser, 2009). Although some researchers may be concerned that such procedures could adversely affect participants' engagement, multiple studies have found that explicitly requesting consent for data sharing does not affect participation rates or the quality and richness of the data provided (Campbell, Goodman-Williams, Javorka, et al., 2023; Cummings et al., 2015; VandeVusse et al., 2022).

Participants also expect that if researchers share their data, they will take measures to protect their privacy and confidentiality by thoroughly de-identifying information before release (Kuula, 2011; Mozersky, Parsons, et al., 2020; Yardley et al., 2014). Qualitative studies often explore sensitive topics, and participants emphasized that they could face negative social, economic, legal, and/or health consequences if their data could be linked back to them (Campbell, Goodman-Williams, Engleton, et al., 2023; Kuula, 2011; Mozersky, Parsons, et al., 2020; VandeVusse et al., 2022; Yardley et al., 2014). However, de-identifying narrative data is challenging because there many types of data that could require remediation.[1] Direct identifiers are unique to a person or otherwise provide a link to one's identity (e.g., name, social security number, address; Centers for Disease Control and Prevention [CDC], 2023). Indirect identifiers are data points that when used in conjunction with other available information may identify a person (e.g., neighborhoods, zip code) and/or unusual information within a data set (e.g., an uncommon racial/ethnic identity, extreme age, unusual occupation; CDC, 2023). For guidance on recognizing direct and indirect identifiers, the

Health Insurance Portability and Accountability Act's (HIPAA) Safe Harbor (HSH) model enumerates 18 types of potentially identifiable information that should be remediated in health-care data sets.

The HSH model does not encompass all types of potentially identifiable information in qualitative studies, and other data points may require remediation because of "deductive disclosure" (Tolich, 2004). In networked contexts (families, groups, organizations, communities), participants have shared experiences that are known and recognizable to other people (Ellis, 1995). Dyadic data are nonindependent: The other half of the dyad has direct knowledge of the events being studied and can identify the research participant (Campbell et al., 2019; Finkel et al., 2015; Joel et al., 2018). Thus, data that may seem adequately de-identified to "outsiders" (e.g., researchers) may still be recognizable to "insiders" in the network. Therefore, qualitative researchers must search not only for direct and indirect identifiers but also dyadic identifiers.

Another challenge of sharing dyadic data is that the "other half" may be motivated to find the data and attempt reidentification. For example, Joel et al. (2018) cautioned that in relationship science, "it is plausible that some will go looking for their partners' responses, given that many romantic partners are indeed motivated to snoop into each other's private information" (p. 87; see also Finkel et al., 2015). Likewise, Campbell et al. (2019) noted that in gender-based violence research, "the details of the acts being studied are indeed known to someone else, someone who knows the study participant, and has already engaged in destructive behaviors toward that person" (p. 4782). Perpetrators often stalk their victims and seek new ways to control and abuse them (Logan & Cole, 2011; Stark, 2009), which could include seeking access to research data. Given these risks, researchers need to consider carefully where they archive dyadic data and how "open" they will make these data to others. Although some proponents of open science advocate for truly open public access to data (for a review, see Siegel et al., 2021), in qualitative research, particularly on sensitive topics, researchers may need to work with archives that vet requests to access data. For example, the Inter-university Consortium for Political and Social Research (ICPSR) was established in 1962 and maintains multiple national archives, including the National Archive of Criminal Justice Data. ICPSR is staffed with experienced curators who can advise researchers and evaluate de-identification coding. ICPSR can also enforce restricted access depending on the sensitivity of the data sets. Likewise, the Qualitative Data Repository was created in 2014 to promote secondary data analyses (see Kapiszewski & Karcher, 2021). This archive provides a curation handbook to guide researchers through the process of preparing and de-identifying data (Demgenski et al., 2021), which are also protected by an application and review process. Unfortunately, it cannot be assumed that the risks of reidentification can be fully mitigated by limiting data access. Furthermore, the presumption of a fully benevolent research community may not always be warranted: Researchers know how to access data, and some individuals may do so for unethical reasons. Given these risks, qualitative data may require extensive remediation to prevent reidentification—regardless of where they are ultimately archived—and researchers need detailed methodological guidance for this work.

Methodological Challenges of De-Identifying Qualitative Data

The types of identifiable information that may require remediation are expansive—direct, indirect, and dyadic. But how will researchers recognize a potentially identifiable data point among the hundreds or thousands of pieces of information that are shared in a qualitative study? One strategy is to conduct a manual, rules-based review of the data whereby researchers develop and follow a codebook that defines specific information (e.g., HSH identifiers), topics (e.g., dyadic events), patterns (i.e., combinations of answers), and/or other features that must be flagged and examined (Walsh et al., 2018). Manual reviews may be a good choice for small data sets with many potential indirect and dyadic identifiers. Another strategy is to use computer automation technology to scan for named entities, such as person, location, dates, and times, which may be particularly helpful when de-identifying large data sets (e.g., unstructured, clinical health data; Walsh et al., 2018). Initial efficacy studies suggested that automated methods did not have optimal detection rates (Kleinberg et al., 2017; Walsh et al., 2018), but in a more recent project, Gupta et al. (2021) developed a natural-language-processing-based de-identification pipeline for large-scale health-care data sets that had far better accuracy (F-1 scores > .90). However, because de-identification software may remove too much or too little data, Gupta et al. noted that "automated de-identification . . . [cannot] replace the need for careful attention from a highly trained human user" (p. 8).

Once a piece of information has been flagged as potentially identifiable, techniques used in quantitative research can be adapted to remediate narrative data (Demgenski et al., 2021; DuBois et al., 2018; Joel et al., 2018; Tsai et al., 2016). For example, blurring data removes some precision and detail (which could lead to reidentification) but retains some information for secondary analysts (Levenstein & Lyle, 2018). Blurring

qualitative data involves replacing original text with altered text [in brackets] that is less specific/more generalized (DuBois et al., 2018). Numerical information (e.g., ages, dates) could be blurred from an exact data point to an interval: age 22 → age [18–25]; year 2008 → year [2000–2010]. Nonnumerical information could be blurred by replacing text with a more generalized term, such as a superordinate category: bipolar disorder → [mental-health condition]; diabetes → [physical-health condition]. Tsai et al. (2016) offered a useful example of blurring individual words and phrases. For example, the original text,

> Just the other day he got angry with me because there was no water and our eldest went to school in a soiled uniform. He threw the empty jerricans at me and you now see the bruise on my left eye (p. 194),

was blurred to: "He got angry with me because there was no water [ ]. He [attacked me] and you now see [my face]." Blurring tries to preserve as much detail and context as possible while acknowledging that the remediation could decrease the usability of the data.

In some contexts, it may not be possible to blur data because even generalized recoding could still provide enough information for reidentification. In such situations, it may be necessary to redact the text entirely (DuBois et al., 2018; Tsai et al., 2016). Redaction may be used for individual words (see Tsai et al., 2016, example above), answers to specific question, a section of an interview (e.g., all questions that describe dyadic events), or a content theme (e.g., all answers pertaining to a specific topic regardless of where in the transcript they appear). As noted previously, it is common practice to use brackets to denote altered text, whereby empty brackets signal redacted information, and researchers could note within the bracket the nature of the information redacted: [details of abuse redacted] (Demgenski et al., 2021). Tsai et al. (2016) emphasized that secondary users should have no expectation of a complete, unredacted transcript because some information simply cannot be shared.

The Current Study

At this juncture in the history of the open-science movement, methodological guidance for sharing narrative data is a work in progress. The Qualitative Data Repository developed general guidelines for de-identifying and archiving narrative data (Demgenski et al., 2021), and some scholars have shared case examples that pull back the curtain to show how this work was done in specific studies (Joel et al., 2018; Tsai et al., 2016). However, there is a pressing need for a deeper exploration of how to remediate nonindependent/dyadic data, particularly in studies on sensitive topics. To that end, we present our methods for de-identifying a set of 32 interview transcripts detailing sexual-assault survivors' experiences with the criminal legal system (Campbell et al., 2022). The interviews were semistructured, so some common topics were explored in all interviews (e.g., the assault itself, initial experiences reporting to the police, later experiences with prosecutors, social support throughout the entire process), but survivors highlighted issues that were salient to their specific mental health and healing journey. During the informed-consent process, we told all potential participants that we were required by the study's funder to archive de-identified interview transcripts in a national research archive. We disclosed this information before participants gave their consent so they could choose to opt out of the study (none did) and before they answered any questions so they could make an informed decision about what they chose to disclose during the interview. As part of the informed-consent process, we also explained how we would de-identify the data before archiving:

> As part of an agreement with our funder, we are required to provide a copy of anonymous interview transcripts (but not the audio recordings) to them for their records. Your safety and confidentiality is very important to us so we will remove/redact the following information from the interview transcript:
>
> - Your name
> - Anyone else who may be identified in the interview
> - Dates
> - Any other details about your case that would be identifiable (i.e., location)
>
> We're required to share the transcripts for our grant, so that other researchers can learn from what you tell us. We will check in with you at the end of the interview, so if there are specific sections or things you've talked about that you don't want included in the transcripts we share, we'll make sure to remove those, too.

These procedures prompted us to ask participants at the end of the interview if they wanted any information removed from their transcripts; two survivors requested specific redactions, neither of which were substantive (Campbell et al., 2022).

At the end of the project, we were responsible for preparing the data for archiving and protecting participants' privacy as promised during the informed-consent process, which was more complicated than we expected.

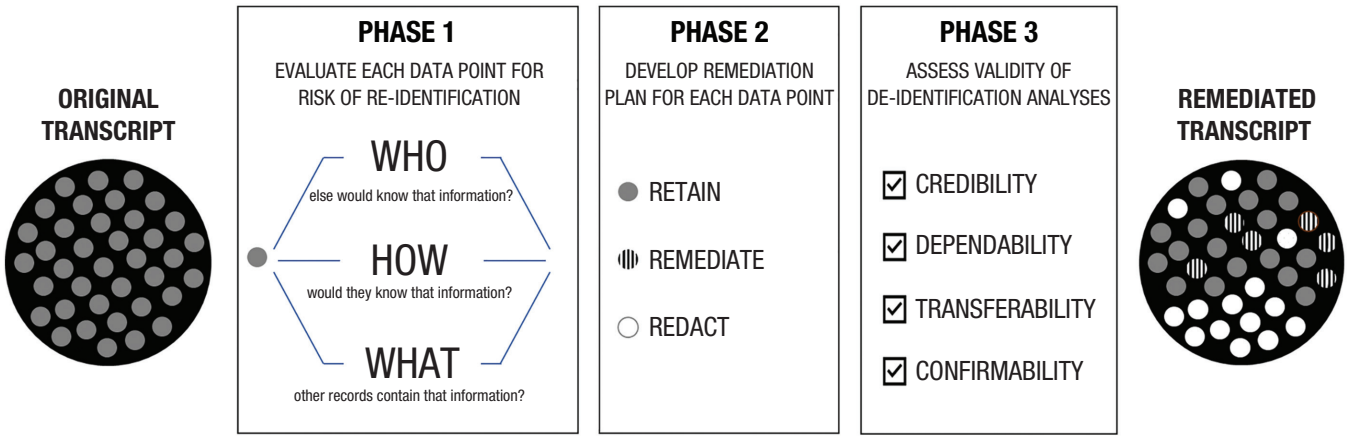

Fig. 1. Overview of a process for de-identifying qualitative data. A qualitative transcript contains many individual data points, represented as individual dots in this figure. Each of those data points must be evaluated for their risk of reidentification and remediated if necessary.

Given the complexity of the data, we used a manual, rules-based approach whereby our team collectively created and implemented a remediation codebook. As we continually revised our processes, the principal investigator (PI) documented the team's work in daily/weekly ethnographic field notes (Emerson et al., 2011), and the research assistants developed memos and audit trails to capture the evolution of the work (Miles et al., 2020). In other words, we approached de-identification as a type of qualitative analysis in its own right. We had used Miles et al.'s (2020) model as our guiding analytic framework for our substantive analyses, so our research team was well versed in developing coding procedures (Phase 1), examining patterns within and across cases (Phase 2), and verifying results and conclusions (Phase 3). We reasoned that we could use a similar approach to de-identify our data: develop a process to distinguish which data may require remediation (Phase 1), decide how to remediate those data (Phase 2), and assess the validity of our de-identification analyses (Phase 3; see Fig. 1). Below, we describe these three phases and then offer case examples of the challenges we encountered preparing our data for archiving.

Phase 1: Developing a Process to Distinguish Potentially Identifiable Data

In Phase 1, our goal was to develop a process that would help us recognize potentially identifiable information within each research transcript (see Table 1). To that end, our task in this phase was to create a coding framework that would help us identify that information. We began by consulting with multiple sources, including the HSH model, federal Institutional Review Board (IRB) regulations, local IRB guidance, and the archiving guidelines of our mandated repository (i.e., ICPSR's National Archive of Criminal Justice Data). We also reviewed the open-science literature and published examples of de-identifying qualitative research (e.g., DuBois et al., 2018; Joel et al., 2018; Tsai et al., 2016). From these sources, we compiled an initial list of types of data that would require remediation (e.g., name, date, locations, dyadic events).

As noted previously, qualitative research is often conducted on sensitive topics with marginalized, minoritized, and/or traumatized populations, so it is also important to consult with the participants themselves. For this study, we asked sexual-assault survivors at the end of the interview what information they wanted removed from their transcript before archiving. Survivors rarely asked for specific redactions, likely because we had already promised them in the informed-consent process that we would remove "other identifiable information" before archiving. To do that work correctly and ethically, we felt we needed additional consultation from subject-matter experts. We reached out to the sexual assault victim service agency that was our collaborative partner in the broader project, prosecutors from the county prosecutor's office that tried these criminal cases, and attorneys from the state's professional organization of prosecuting attorneys. These colleagues reminded us that because our research focus had been sexual-assault survivors whose cases had been prosecuted by the legal system, all survivors had already "gone public" with details of the assault during trials or plea hearings. This context did not absolve us of our responsibility to protect the research data (because many details were already part of a public record); rather, this meant we needed to take *extra* precautions so that our research-interview transcripts could not be linked to trial transcripts, which identify the survivor and perpetrator(s) by name.

These consultations prompted us to consider the extent to which data in research transcripts are also contained in public records. The Freedom of Information

Table 1. Three-Phase Process for De-Identifying Qualitative Data

		Actions to complete tasks		
Phase goal	Tasks	Actions	Examples	Specific examples from this project
Phase 1: develop a process to distinguish potentially identifiable data	Create a coding framework	Consult with stakeholders	Regulatory guidance	Federal Health Insurance Portability and Accountability Act (HIPAA), federal and local Institutional Review Board (IRB) guidance, archiving guidelines from the National Archive of Criminal Justice Data (NACJD)
			Open-science literature on de-identifying qualitative research	DuBois et al. (2018), Joel et al. (2018), Tsai et al. (2016)
			Research participants	Sexual-assault survivors
			Subject-matter experts familiar with the population studied	Victim service agency staff, prosecutors
			Publicly available records that may contain same/similar information as the research-interview transcripts	Court transcripts
		Draft a codebook	Named entities	Names, dates, locations
			Identifiable topics	Details of the sexual assault, postassault help-seeking experiences, criminal trial proceedings, survivors' health and healing
			Guidance for evaluating ambiguous information	Who would know this information? How would they know this information? What other records contain this information?
Phase 2: remediate potentially identifiable data	Form coding team	Hire and train coders	Include coders with varying levels of familiarity with the data	Original project interviewers and another sexual-violence researcher, unaffiliated with the original project
	Review transcripts and propose remediation plans	Highlight each data point and create an audit trail containing proposed edits	Draft blurred text	Converted numerical data to an interval range, replaced labeled terms with superordinate categories, or altered specific text with more generalized text
			Bracket redacted text	Removed sections of text; when feasible, inserted a summary written by the research team that contains information on key variables
			Review proposed remediation plans and discuss as a team	Two supervisors reviewed each transcript to verify whether other data points required remediation.
	Implement remediation plans	Edit and redact text	Insert blurred text, remove redacted text, and insert summaries of large redactions	Two supervisors verified whether remediation plans were implemented accurately and then reread the de-identified transcripts to recheck whether other data points required remediation.
	Provide support to coding team if there will be repeated exposure to traumatic content	Check in with team members regarding their experiences of vicarious trauma	Provide information and support regarding the emotional impact of repeated exposure to traumatic content and offer supportive resources	Principal investigator and supervisors checked on the emotional well-being of team members throughout this phase of the project and offered tailored supportive measures.

(continued)

Table 1. *(continued)*

Phase 3: assess the validity of the de-identification analyses	Select validity standards (e.g., Lincoln & Guba, 1985)	Assess credibility	Use of prolonged engagement, persistent engagement, and member checks to assess accuracy of the findings	Engaged with this community and victim service organization for 10+ years; conducted member checks with victim advocates
		Assess dependability	Use of codebooks, memos, and audit trails to document analyses	Created codebook to document process and audit trails for each data point remediated
		Assess transferability	Provide audiences with sufficient detail about the project so they can assess whether conclusions are transferable to other settings	Outlined a three-phase process for de-identification as a coding framework for other projects
		Assess confirmability	Consider how researchers' positionalities affected the processes and findings and when necessary, recenter the participants' views	Developed positionality statements and discussed how to create remediation plans that balanced participants' wishes to contribute to science (through usable data) and protecting safety, privacy, and confidentiality

Act (FOIA) makes a great deal of information available to the public, and depending on how those records are/are not redacted, these other sources could be leveraged to reidentify a research transcript. Thus, researchers may need to "consult" with those records and cross-check overlapping information. In our project, the state prosecutor's association advised us to request trial transcripts through FOIA to understand what information is contained in those records to guide our de-identification protocol. We obtained four trial transcripts and cross-checked them against research transcripts. From these sources, we created a list of topics, questions, and themes that would likely need remediation (e.g., direct- and cross-examination questions during the sexual-assault trials/plea hearings).

Our next action was to draft a codebook that listed types of information that might be identifiable, per our consultations. However, as we built this codebook, we recognized it was nearly impossible to preidentify all possible data points that may require remediation, so we developed three guiding questions to help us evaluate ambiguous information: (a) Who else would know that information (e.g., the perpetrator, family/friends, service providers, prosecutors)? (b) How would they know that information (e.g., because they directly committed the act/event, because they witnessed an event, because they were directly told this information)? and (c) What other records contain that information (e.g., a court transcript, police reports)? For this third guiding question, we realized that the research-interview transcript should also be considered a record that could provide reidentification clues: Multiple data points—when reviewed together as a set—could inadvertently reveal information. For example, we might think that data point "X" required no remediation, but later in the transcript, that unredacted information could help the reader reidentify data point "Y." In other words, we would need to imagine someone reading ahead or backtracking through material in the transcript and consider how unredacted details could be used in combination to provide reidentification clues. Thus, our coding processes needed to consider how remediation decisions knit together to create an overall document that adequately protected participants' identity, privacy, and confidentiality.

Phase 2: Remediating Potentially Identifiable Data

In Phase 2, our goal was to use this coding process to identify and remediate potentially identifiable information (see Table 1). Our first task was to form a coding team so that each task was cross-checked and verified by multiple analysts (see MacQueen et al., 2008). We hired coders with varying degrees of familiarity with these data, including the original interview staff who worked closely with the community and the participants, which could help them recognize potential deductive disclosures. We also hired one new staff member, someone who also conducts sexual-violence research but was

not involved in our original project. We reasoned that including someone who had not heard the survivors' stories firsthand could help us evaluate whether specific remediations and redactions would still leave the data interpretable to others. Two coders (an original interviewer and the new team member) then reviewed the transcripts and tagged potentially identifiable text. We had used Atlas.ti (Version 8) for our substantive analyses but switched to Microsoft Word for our de-identification project, assuming it would be easier to make remediations in word-processing software. However, we found that many remediations required multiple rounds of team review and debate, which needed to be captured in an analysis audit trail. Each data point to be remediated was highlighted with a Microsoft Word comment box, which became our audit trail as we considered remediation options. Coders proposed edits to the text in the comment box; transcripts were not altered in this phase. Our first-line option was to try blurring the data by (a) converting numerical data to an interval range, (b) replacing labels with superordinate categories, or (c) altering specific words with more generalized text. Our second-line option was redaction, which we considered only after we determined that blurring would leave the data too identifiable (for how we conducted this work and tried to balance privacy and utility to secondary data users, see case examples below).

After the coders developed draft remediation plans, the transcripts were reviewed by two supervisors to check whether other data points needed remediation, consider how information not tagged for remediation could potentially reidentify other details in the interview, evaluate whether the proposed remediation plans were sufficient, verify whether we were using consistent remediation strategies across cases, and elevate complex challenges for full-team discussion. Once we had what we believed to be the final remediation in each data point's comment box (i.e., the agreed-on blurred text or redactions), the coders saved the transcript under a new name and date and then implemented the remediation plan in the text of the transcript. The comment box was not deleted in this version so that a supervisor could recheck whether the text had been remediated correctly per team consensus. Any errors were reviewed and addressed between the coder/supervisor as needed. The file was resaved under another new file name and date, and then a supervisor rechecked the remediated text for accuracy and finally, deleted all comment boxes. The cleaned, de-identified transcripts were reviewed again to assess whether any other data points emerged within or across transcripts that were potentially identifiable (if so, we repeated team review and consensus). This process dictates many reviews of the data, which in our project repeatedly exposed team members to traumatic content. We encourage research teams to be aware of the impact of vicarious trauma and take proactive steps to support staff while conducting this phase (see Campbell, 2002, 2023).

Phase 3: Assessing the Validity of the De-Identification Analyses

In Phase 3, our goal was to assess the validity of our de-identification methods (see Table 1). As noted above, de-identification was a more deliberative decision-making process than we had expected, so we felt it was necessary to conduct a formal validity assessment. Numerous strategies have been proposed for assessing validity in qualitative research (Creswell & Poth, 2018), and we drew on the classic four standards outlined by Lincoln and Guba (1985).[2] To establish credibility (i.e., confidence in the accuracy of the findings), we had prolonged engagement and persistent observation in this community (10+ years), which was helpful in recognizing and remediating identifiable locations and local history. Another common strategy to verify credibility is to conduct member checks whereby participants or other stakeholders are asked to review and critique analyses, interpretations, and/or conclusions. In the context of de-identification analyses, researchers could ask the participants themselves to review remediated transcripts and verify whether they felt their identity was sufficiently protected. In this project, we were not able to do member checks with our primary research population (sexual-assault survivors) for two reasons. First, our consent process framed participation as a one-time interview, with no open door for later engagement. This is a challenging population to recruit (see Campbell et al., 2022), and our partner victim-advocacy organization strongly recommended we request only a single interview/interaction to boost participation rates. Second, member checking with survivors would require them to reread and revisit traumatic material, which some may understandably prefer not to do. However, we did engage victim service agency advocates for member checking our substantive analyses (see Campbell et al., 2022) and these de-identification analyses. We provided a set of de-identified transcripts and asked staff whether they recognized the survivors in each case and/or whether there was additional information that should be redacted. In all cases, staff indicated that they could not identify the survivor in each interview, and they did not flag additional content for remediation. Given that the advocates had worked with these survivors as clients for years, we felt reasonably assured we had created and implemented a protocol that successfully de-identified the transcripts.

To establish dependability (i.e., the findings are consistent and could be repeated), we maintained an audit trail throughout the de-identification analyses that

tracked coding processes, coding decisions, remediation plans, and remediation implementation (see summary in Table 1). Regarding transferability (i.e., the findings have applicability in other contexts), Lincoln and Guba (1985) argued that researchers need to provide their audiences with sufficient detail about what happened in the setting of interest so that they can assess whether the conclusions are transferable to other settings. The specific details that needed to be redacted in our study are unique to this project, but we offer our process (see Table 1) and our three guiding questions (Who else would know that information? How would they know that information? and What other records contain that information?) as a transferable set of focal concerns that researchers must attend to when de-identifying qualitative data. To establish confirmability (i.e., the findings reflect the participants' views, not the researchers' biases), the PI kept field notes, and the PI and primary analysts wrote reflexive memos throughout the project to ensure that the findings did not reflect the biases of the research team (for how we engaged our positionalities throughout the study, see Campbell, Goodman-Williams, Engleton, et al., 2023). For these de-identification analyses, we wrestled with the fact that our lived experiences are strikingly different from those of our research participants, and thus, we felt we could not fully understand what might compromise their safety, privacy, and confidentiality. We found Wood's (2015) guidance helpful because she emphasized that survivors are the experts of their own lives and that engaging their voices and preferences is critical for trauma-informed research. We had asked survivors what they wanted removed from their transcripts, and as we evaluated the remaining content, we tried to leverage our experiences working with survivors for decades to make decisions that would protect their identities.

De-Identification Case Examples

It is challenging to "show" our work because we cannot provide original text passages side by side with remediated text. However, we can provide examples of common problems we encountered as we removed "name, date, location, and any other identifying information" from the transcripts.

Names

In these interviews, survivors sometimes mentioned their family, friends, partners, service providers (e.g., advocate, detectives, prosecutors), and or the assailant(s) by name. In all instances, we removed the name and replaced it with a label denoting the person's relationship to the survivor (i.e., blurring by superordinate category). However, in the case of family members, we felt it was prudent to not specify exact relationships and simply label individuals as "[family member]." For example, some survivors discussed that when they told their mothers, siblings, or partners that they had been assaulted, those individuals disclosed in turn that they too had been sexually victimized in their lives. If our labels included the specific family relationship (e.g., "mother"), then these secondary disclosures of assault would have been identifiable. To protect their privacy, we assigned the generic label "[family member]," which obscures information secondary researchers may want regarding social-support experiences, but we prioritized the privacy of those individuals when making remediation decisions. Likewise, some survivors described how the criminal trial put considerable strain on family relationships, and the specifics of who was/was not a trusted confidant may have been known by others (who) because there may have been direct conflict (how). Using a blurred superordinate category ([family member]) provided more protection to all individuals and more context for secondary users than would even higher superordinate categories (e.g., [support person]).

Dates/passage of time

The survivors we interviewed experienced a significant delay between when they reported the assault to the police and when their case was ultimately prosecuted. The police had closed their cases prematurely before completing an investigation and testing all available forensic evidence (e.g., a rape kit/sexual-assault kit containing blood, semen, saliva, hair sample; see Campbell et al., 2022). Given this context, the years when events occurred and how long survivors waited for justice was a salient theme in the interviews, but this information was highly identifiable (e.g., who: many people knew that information; how: through direct and indirect mechanisms; and what: information was contained in multiple records). Some survivors provided this information directly (i.e., participants gave dates, years, length of time in between), but some marked time by their children's ages or other significant events in their lives (e.g., "that happened when my son was 3"; "it happened after my mom died"). Thus, we had identifiable numerical and social/contextual markers of the passage of time.

To strip all date-related information would render the transcript significantly less usable for secondary analysis. Therefore, we removed specific dates/years/ages, and whenever possible, if we had enough information to compute a span of time, we inserted that blurred information. For example, if survivors indicated that they reported the assault to the police in 1990 but the case was closed and then they were recontacted in 2012 when their untested rape kit was located and sent for testing, we removed both dates and inserted "[survivor contacted

22 years later]." If a survivors referenced that an event happened when their son was 3 years old, we removed that text and replaced it with blurred information about how much time had passed, if we could compute it; otherwise, the text was redacted. Likewise, survivors typically provided specific information about the assailants' conviction sentences, which included what appeared to be exact statements from judges made during sentencing hearings (which is identifiable by multiple people [who], means [how], and sources [what]). We blurred the criminal sentences to numerical ranges (e.g., 5 years was coded as 2–5 years), but it was not uncommon that the sentence itself was a range. If a judge sentenced the assailant to 5 to 20 years, it meant that the rapist would serve a minimum of 5 years and a maximum of 20 years, so we blurred both ends of that range: "[minimum of 2–5 years; maximum of 11–20 years]."

Location

The location where this study took place was required information in the grant application and final grant report, both of which are publicly available documents. State and national media have reported on this city and its long-standing problems with untested sexual-assault evidence, so there are multiple ways secondary users can learn the name of the city where we conducted this study. Nevertheless, we redacted the name of the city from the transcripts, per IRB guidance. In the event survivors provided addresses or neighborhood locations, we redacted that information (because it would be identifiable by multiple people [who], means [how], and sources [what]). We evaluated whether we could blur that information by replacing it with census-track data or zip-code data, which could be useful to secondary users interested in geographic patterns in crime. However, we felt the safety risk of providing any geographic information about survivors' location was simply too high, so we remediated the data by redaction. We also had to leverage our insider knowledge of this city to recognize local nomenclature that revealed locations and neighborhoods, and that text was also redacted.

Other identifiable information: the assault

Perhaps the most significant challenge we faced was how to handle the narrative material regarding the sexual assault itself, which was a substantial part of the interview, and details of that experience were often retold in response to other questions. This is highly identifiable dyadic data, known to multiple people, including the assailant, and the details of that event are captured in the court transcript. We tried to blur these data because wholesale removal of this material would significantly alter the usability of the transcripts for secondary analysis, but there were numerous individual words and sentences that needed to be changed. In one of our team meetings, a coder noted, "I feel like I'm rewriting her story and that doesn't feel ok" (quote recorded in PI's field notes). That comment prompted us to try a different approach: We would redact the actual text (i.e., survivors' words) but include a brief summary of the assault written by the researchers, which clearly delineates survivors' words from our interpretation/understanding of their words. We believe that survivors' stories are their stories and that altering their words may feel disrespectful and disempowering. Replacing their words with ours may likewise feel disrespectful and disempowering, but it also signals our assumed responsibility for these redactions and any critiques of these decisions.

We debated how much detail could be safely shared in these assault summaries that we would be drafting. Our aim was to provide key variables that would likely be of interest to secondary users without the details or context that would identify the assault/case/survivor. For example, in our field of study, it is common to examine how postassault help-seeking and health outcomes vary as a function of assault characteristics (e.g., whether assault was perpetrated by a stranger, whether there was physical violence, whether a weapon was used). As we removed the narrative of the assault, we tried to provide "answers" on key variables in ways that would not reidentify the case. For example, here is the redacted text from one transcript: "It was in [REDACTED: Date] I [was] [REDACTED: kidnapped and raped at gunpoint by multiple assailants, stranger perpetrated assaults]." We discussed at length whether this summary was also too identifying (i.e., a kidnap and rape at gunpoint) because in some studies, this would be an uncommon assault profile (i.e., an indirect identifier; CDC, 2023). Unfortunately, in this jurisdiction, it was not. These details do not uniquely distinguish this case because many survivors we interviewed were assaulted in similar ways. By contrast, we found that for assaults committed by individuals known to the survivors (e.g., partners, friends, family members), we needed to be more restrictive about what was shared, as these two examples illustrate:

Interviewer: Will you tell me about what happened with the assault?

Participant: [REDACTED: intimate partner perpetrated assault]

Interviewer: Will you tell me about what happened in the assault?

Participant: [REDACTED: assaulted by multiple assailants; one was a friend, the rest were strangers]

These excerpts highlight how researchers may need to consider the potential identifiability of information both within and across cases when remediating data because what is unique or common in a data set will vary.

Other identifiable information: postassault help-seeking experiences

The interview also explored survivors' initial help-seeking experiences reporting the rape to the police. We anticipated that these sections of the transcript may also need complete redaction because the information therein was also dyadic (i.e., between the survivor and the police/detective, thus known to multiple people through multiple means). When we reviewed the sample court transcripts, however, there was not much information about survivors' initial reporting experience in those records, and what was included tended to be factual events (e.g., the date the report was made, which could be resolved through blurring; see above). Thus, we were less concerned that information in the research transcript could be linked to information in a court transcript. As we studied these sections of the interview closely, we also noticed remarkable similarity within and across cases. In other words, the ways in which police treated victims was strikingly repetitive—repetitively negative and victim blaming—which was a key substantive finding in the study. In a team discussion (captured in the PI's field notes), a coder said, "The police aren't very original in how they do this." When we read these sections of the transcripts back-to-back, the lack of unique detail was noticeable, which allowed us to preserve more of the data, as this example illustrates:

> Well firstly, I have two separate incidents in the rape kit backlog. [REDACTED: Details of first assault by known assailant.] They assaulted me I was [REDACTED: age at time of assault: 18–24] at the time, in [REDACTED: Date]. I reported it to the police, and outside of treating me basically like garbage, and like a whore, and like a liar, they threw me away. And so in [REDACTED: Date] when I was kidnapped off the street by a stranger, who turned out to be a serial rapist, I was a little more hesitant to call the police.

We did not need to blur the specific words "garbage," "whore," or "liar" because these were common terms used by survivors when they recounted how they were treated by the police. Likewise, many of the women we interviewed were abducted by strangers, and this text required no additional remediation.

Other identifiable information: trial experiences

The survivors' descriptions of their court experiences included both event-level information (which is dyadic and known to multiple sources, through multiple means, recorded in official records) and their internal reflections about those events. The specific questions asked at trial are also captured in the court transcript, so those sections of the research transcript needed remediation. For instance, when survivors described their trial experiences, they recounted what appeared to be dialogue (e.g., "I said . . . , he said . . . "). It was often unclear whether these were actual statements (which would be in the transcripts) or whether survivors were relaying information by using "I said . . . " as a narrative device. Given this ambiguity, we typically chose to redact the text in case it was a direct quote. Defense attorneys used common strategies to try to discredit victims, so many details about survivors' court experiences were not as identifiable as we anticipated, as this excerpt illustrates:

> I was able to let his lawyers know, [REDACTED: participant quote from court]. And that was the hardest thing I had to do. It was so hard because I broke down on a stand and cried. It was so hard because for . . . lawyer to sit there and make it seem like it was my fault, I wanted it. And I knew you, and you know for a fact, I didn't know. You knew this, you knew it. So when . . . [lawyer] saying that hurt my feelings.

We were surprised by the extent to which blurring was a feasible strategy in these interview sections, and we were able retain largely coherent narratives of the trial events.

Other identifiable information: health and healing

Throughout the interviews, survivors shared details about their lives and healing journeys. Many described the short-term and long-term impact this victimization had on their health and well-being, including vivid, powerful stories about their mental health, physical health, disabilities, pregnancies, and substance use. Some of this information may be known to others (who), through multiple means (how), and documented in various medial/social-service records (what). Even though access to those records is restricted by HIPAA, we were mindful that if the research transcript and the court transcript were eventually linked, survivors' names would be discoverable, and thus, this private information (which was

not shared in court) about their health and healing would be known. In addition, some participants described having multiple, relatively low-prevalence, co-occurring health conditions. We were concerned that this kind of health information that is known to others (who) through multiple means (how) could itself be identifying in combination with other details from the transcript if those others gained access to the research transcript. To address these concerns, we used blurring when possible, such as inserting subordinate categories (e.g., "[health condition]") that provided virtually no information about the nature of the health matter. We did not feel it was appropriate for the research team to write a summary about these experiences to include in place of redactions because, again, it was private health information. Thus, some of the deepest, richest data were redacted in the edited transcripts for secondary analysis.

Discussion

A fully open approach to psychological research will be challenging given the diversity of epistemologies, methodologies, and areas of inquiry that coexist under its disciplinary tent. Releasing all data and all research materials to all audiences may not be reasonable and responsible practice—and to be clear, few proponents of open science advocate for such expansive strategies (for a review, see Siegel et al., 2021). Researchers can make psychological research more transparent and accessible, and how this is done must be tailored to the unique contexts of a study's methods, questions, and research population. The focal concerns in qualitative inquiry are different than quantitative research, particularly the challenges of de-identifying narrative information so that sharing data does not lead to participant reidentification. Qualitative instructional texts typically do not address open-science practices or provide recommended methods for remediating narrative data (e.g., well-regarded texts by Braun & Clarke, 2022; Corbin & Strauss, 2014; Creswell & Poth, 2018; Denzin & Lincoln, 2018; Miles et al., 2020; Patton, 2015; Saldaña, 2021; Tracy, 2019, do not address these issues). To address this gap, we developed a framework for de-identifying narrative data, provided case examples of how we prepared our data for archiving, and proposed strategies for assessing the validity of this type of coding.

This framework prompts researchers to consult diverse sources and stakeholders to understand what information may be identifiable within the context of a specific study. In this project, we followed federal and local IRB regulations, but we found ourselves wanting and needing more consultative support from our IRB on how to remediate our data than they were able to provide. To be clear, we are not disparaging our IRB colleagues but, rather, underscoring the limits of their preparation and training and, thus, the extent to which they can advise researchers on these matters. IRB members typically do not receive training on de-identification methods for either quantitative or qualitative data (Meyer, 2018; Mozersky, Walsh, et al., 2020). For example, Mozersky, Walsh, et al. (2020) interviewed qualitative researchers, data archivists/curators, and IRB members about their knowledge and experience with open-science practices with narrative data. They found that "IRB members and data curators are not prepared to advise researchers on legal and regulatory matters, potentially leaving researchers who have the least knowledge with no guidance" (p. 1). Likewise, Meyer (2018) noted, "In the short term, institutional privacy offices will tend to have more expertise in recognizing re-identification risks and in recommending solutions than will most IRBs" (p. 135).

We add our voice to other scholars who have called for more training for IRB staff on open-science practices and a larger role for IRBs in the oversight of data-dissemination protocols (e.g., Meyer, 2018; Mozersky, Walsh, et al., 2020). Although multiple U.S. research funding agencies, including the National Institutes of Health and the National Science Foundation, have policies requiring data sharing, to date, there has been minimal U.S. federal guidance on whether and how IRBs should monitor data-sharing practices (Meyer, 2018; Mozersky, Walsh, et al., 2020). Thus, most IRBs do not require researchers to submit their data-dissemination protocols for approval (Meyer, 2018). Institutions could implement local policies that require researchers to specify these plans as part of their IRB approval process, even if federal Common Rule guidance does not yet require such oversight (Meyer, 2018). Likewise, professional organizations could append their ethical-standards documents to offer recommendations for protecting privacy when sharing data. As Siegel et al. (2021) noted, mandates to comply with open-science practices are outpacing the development of nuanced recommendations for how to do this work safely. For qualitative scholars, this liminal space is challenging to navigate, particularly in projects in which reidentification could pose serious health and/or legal consequences for participants.

Given these risks, researchers should also consult with relevant subject-matter experts to understand the potential identifiability risks of their data. At a minimum, researchers should seek informed consent from their participants to share their data with others, and we contend that it is also important to ask whether participants want specific information withheld. In this study, when survivors were given that option, they did not remove any substantive data. Wood (2015) emphasized that survivors of gender-based violence are well positioned to evaluate threats to their safety, and here we extend Wood's argument to prompt researchers to check with participants

about their safety concerns when de-identifying data. We also recommend consultation with other subject-matter experts who may be able to envision risk scenarios that neither the research team nor participants considered. For example, the subject-matter experts we consulted (victim advocates and attorneys) challenged us to think about how publicly available information (through FOIA and/or court records) may complicate de-identification. Because more information is available through public means, researchers will need to consider how the data sets they share can be used in conjunction with other records to reidentify data. The HSH model may not provide sufficient coverage, so we developed three guiding questions researchers should consider for each data point that may be reidentifiable. First, who else knows that information—and what is known about their intentions, motivations, and behaviors? Second, how do they know that information—and what about that context gives them unique information? Third, what other records contain this information—and how could the information be cross-linked across records? These questions require researchers to evaluate risk at both a macro level and a micro level, data point by data point, which transforms de-identification from a simple task to a thoughtful process.

The process of de-identifying data should be tailored to the type of data and the relative risks of reidentification. At a minimum, we recommend that researchers consider each of these three focal questions, but depending on the answers, the steps we outlined here could be simplified. For example, if there is no other "who" with knowledge about the issues studied (the first question), then the HSH model may be sufficient. Likewise, depending on how others might have identifiable information (the second question) or what other records contain reidentification clues (the third question), researchers might make different remediation decisions. For instance, the location in which a study is conducted is a key factor identifying the population from which a study's sample is drawn—and if that information can be redacted and is otherwise unknowable, then other details might be "spared" from remediation. In this project, we were mindful that the name of the city in which this study was conducted is available through multiple public records.[3] However, if we could have protected that information, we could have retained more context about survivors' family relationships, their help-seeking experiences, and their health and healing journeys because there were no other clues that would have narrowed the population pool so markedly. In some projects, it may be possible to evaluate trade-offs of retaining some types of information versus others, but in our study, we had to be careful and conservative because a highly identifiable data point could be discovered through public records.

Given these constraints, de-identifying our data was time-consuming and stressful (for a firsthand account for how difficult this work is for a research team, see also Joel et al., 2018). This was the last task in closing out a federal grant, which had already required years of stakeholder engagement, interviewing survivors, coding and analyzing traumatic data, and preparing a lengthy substantive final report. Preparing the data for archiving was an allowable grant expense, so we had budgeted time (3 months) and resources (four half-time staff members: two frontline coders and two supervisors), but we underestimated both. The primary work was completed in 4 months, and we needed a fifth staff member to assist as our close-out date loomed. Furthermore, de-identification required repeated, sustained exposure to traumatic material (reading the written transcripts closely multiple times) after everyone on our team had already endured repeated, sustained exposure to traumatic material (conducting the interviews and primary analyses). This is the very definition of vicarious trauma (Cieslak et al., 2014; Lipsky, 2009), which we had been proactively addressing during data collection and analysis (for strategies, see Campbell, 2002, 2023) but struggled to hold at bay during this work. We learned the hard way that what is identifiable and what is upsetting are not necessarily the same thing. The assault narratives were both, and our remediation strategy required close review of that material to prepare the summary redactions. The descriptions of survivors' experiences with the criminal legal system were not as identifiable and could be retained with minimal blurring or redaction, but these sections were painful to review over and over again as we settled on specific word-by-word remediation plans. We felt an ethical and moral obligation to protect these data, which meant that we had to check, recheck, and recheck again—and each check had an impact on us. We share these experiences to highlight that the costs of open-science practices are many and varied and to forewarn other research teams who may also need to de-identify traumatic research about the risks of vicarious trauma and to plan adequate resources, time, and support to do this work carefully and responsibly.

Throughout this project, we wondered whether we should have asked our funder for an exemption to the mandated data-archiving requirement. If we had, it is possible we would have received a waiver given the sensitivity of the data. So why did we press on? First and foremost, because our research participants explicitly said they wanted others to learn from their experiences and they wanted these data to be a catalyst for criminal-justice-system reform (Campbell, Goodman-Williams, Javorka, et al., 2023). Participants wanted to help other sexual-assault survivors, and they felt that

making their data available to researchers would promote education, training, advocacy, and new discoveries (Campbell, Goodman-Williams, Javorka, et al., 2023). There are risks associated with sharing data, but our mandated archive (ICPRS's National Archive of Criminal Justice Data) is a secure resource that provides multiple options for protecting data and restricting access to qualified users (ICPSR, 2023). We also pressed on because this was an opportunity to study the complexities of open-science practices in qualitative research. We concur with other qualitative methodologists who have argued that researchers should not uncritically follow open-science practices that were developed for other types of data but that the field can explore how to promote transparency within this type of inquiry. In this project, we will not be able to provide fully transparent, unredacted narratives, but we can be transparent about the processes and decisions we made to remediate the data. The strategies we employed to protect our participants' privacy fundamentally changed the data that can be shared with others. The archived transcripts are different, both quantitatively (there is less information to share) and qualitatively (there is less richness and detail to share), from the original transcripts. Whether these remediated data will be useful to secondary analysts and yield new discoveries is an open question, but we do hope the methods outlined here will be useful in advancing nuanced approaches to open science.

Transparency

Action Editor: David A. Sbarra
Editor: David A. Sbarra
Author Contributions

Rebecca Campbell: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Writing – original draft.

McKenzie Javorka: Conceptualization; Formal analysis; Methodology; Project administration; Supervision; Writing – review & editing.

Jasmine Engleton: Formal analysis; Writing – review & editing.

Kathryn Fishwick: Formal analysis; Writing – review & editing.

Katie Gregory: Formal analysis; Supervision; Writing – review & editing.

Rachael Goodman-Williams: Formal analysis; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by a grant from the U.S. Department of Justice, Office on Violence Against Women (2018-SI-AX-001).

Open Practices

This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Rebecca Campbell https://orcid.org/0000-0003-0442-9835

Acknowledgments

The opinions or points of view expressed in this document are solely those of the authors and do not reflect the official positions of any participating organization or the U.S. Department of Justice. We assure that no financial interest or benefit has arisen from the direct applications of this research. The analyses presented in this article have not been previously published.

Notes

1. Qualitative data have many forms (e.g., words, images, and sounds), which have distinct identifiability considerations. Open-science practices with qualitative data have focused primarily on de-identification of written interview transcripts (Demgenski et al., 2021; DuBois et al., 2018; Tsai et al., 2016), which is our focus here as well.
2. As the literature on open science in qualitative research evolves, it may be fruitful to explore different approaches for assessing the validity of de-identification coding. In this project, we offer a reinterpretation of Lincoln and Guba's (1985) classic framework as a starting place and acknowledge that model was not originally intended for this kind of analysis.
3. It is beyond the scope of this discussion to delve into why funders may choose to make a study's location accessible information, but we note there is a long tradition of this practice in participatory action research projects with the criminal legal system (see Kennedy, 2012). In fact, such projects are often referred to and become known by the name of the city itself. Such practices can facilitate utilization of a project's findings because it helps practitioners know who/where to reach out for protocols, training programs, and so on (Klofas et al., 2010). Thus, in some contexts, making the location of the study public can have important benefits.

References

American Statistical Association. (2017). *Recommendations to funding agencies for supporting reproducible research*. https://www.amstat.org/docs/default-source/amstat-documents/pol-reproducibleresearchrecommendations.pdf

Bennett, E. A. (2021). Open science from a qualitative, feminist perspective: Epistemological dogmas and a call for critical examination. *Psychology of Women Quarterly*, *45*(4), 448–456. https://doi.org/10.1177/03616843211036460

Brabeck, M. M. (2021). Open science and feminist ethics: Promises and challenges of open access. *Psychology of*

Women Quarterly, *45*(4), 457–474. https://doi.org/10.1177/03616843211030926

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. Sage.

Campbell, R. (2002). *Emotionally involved: The impact of researching rape*. Routledge.

Campbell, R. (2023). Revisiting *Emotionally involved: The impact of researching rape* twenty years (and thousands of stories) later. In M. Horvath & J. Brown (Eds.), *Rape: Challenging contemporary thinking* (2nd ed., pp. 12–27). Routledge. https://doi.org/10.4324/9781003163800-3

Campbell, R., Goodman-Williams, R., Engleton, J., Javorka, M., & Gregory, K. (2023). Open science and data sharing in trauma research: Developing a trauma-informed protocol for archiving sensitive qualitative data. *Psychological Trauma: Theory, Research, Practice and Policy*, *15*(5), 819–828. https://doi.org/10.1037/tra0001358

Campbell, R., Goodman-Williams, R., & Javorka, M. (2019). A trauma-informed approach to sexual violence research ethics and open science. *Journal of Interpersonal Violence*, *34*(23–24), 4765–4793. https://doi.org/10.1177/0886260519871530

Campbell, R., Goodman-Williams, R., Javorka, M., Engleton, J., & Gregory, K. (2023). Understanding sexual assault survivors' perspectives on archiving qualitative data: Implications for feminist approaches to open science. *Psychology of Women Quarterly*, *47*(1), 51–64. https://doi.org/10.1177/03616843221131546

Campbell, R., Gregory, K., Javorka, M., Engleton, J., Goodman-Williams, R., & Fishwick, K. (2022). *Evaluating a victim notification protocols for untested sexual assault kits (SAKs)* (Final Report Award 2018-SI-AX-0001). Office on Violence Against Women. https://doi.org/10.3886/ICPSR38921.v1

Centers for Disease Control and Prevention. (2023). *What is personally identifiable information?* https://www.cdc.gov/nchs/training/confidentiality/training/page581.html

Cieslak, R., Shoji, K., Douglas, A., Melville, E., Luszczynska, A., & Benight, C. C. (2014). A meta-analysis of the relationship between job burnout and secondary traumatic stress among workers with indirect exposure to trauma. *Psychological Services*, *11*(1), 75–86. https://doi.org/10.1037/a0033798

Class, B., de Bruyne, M., Wuillemin, C., Donzé, D., & Claivaz, J. B. (2021). Towards open science for the qualitative researcher: From a positivist to an open interpretation. *International Journal of Qualitative Methods*, *20*. https://doi.org/10.1177/16094069211034641

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.). Sage.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th edition). Sage.

Cummings, J. A., Zagrodney, J. M., & Day, T. E. (2015). Impact of open data policies on consent to participate in human subjects research: Discrepancies between participant action and reported concerns. *PLOS ONE*, *10*(6), Article e0125208. https://doi.org/10.1371/journal.pone.0125208

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin*, *144*(7), 757–777. https://doi.org/10.1037/bul0000154

Demgenski, R., Karcher, S., Kirilova, D., & Weber, N. (2021). Introducing the qualitative data repository's curation handbook. *Journal of eScience Librarianship*, *10*(3), Article 8. https://doi.org/10.7191/jeslib.2021.1207

Denzin, N. K., & Lincoln, Y. S. (2018). The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (5th ed., pp. 1–20). Sage.

DuBois, J. M., Strait, M., & Walsh, H. (2018). Is it time to share qualitative research data? *Qualitative Psychology*, *5*(3), 380–393. https://doi.org/10.1037/qup0000076

Ellis, C. (1995). Emotional and ethical quagmires in returning to the field. *Journal of Contemporary Ethnography*, *24*(1), 68–98. https://doi.org/10.1177/089124195024001004

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing ethnographic fieldnotes*. University of Chicago Press.

Feldman, S., & Shaw, L. (2019). The epistemological and ethical challenges of archiving and sharing qualitative data. *American Behavioral Scientist*, *63*(6), 699–721. https://doi.org/10.1177/0002764218796084

Field, S. M., van Ravenzwaaij, D., Pittelkow, M., Hoek, J. M., & Derksen, M. (2021). *Qualitative open science – Pain points and perspectives*. OSF. https://doi.org/10.31219/osf.io/e3cq4

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*(2), 275–297. https://doi.org/10.1037/pspi0000007

Gupta, A., Lai, A., Mozersky, J., Ma, X., Walsh, H., & DuBois, J. M. (2021). Enabling qualitative research data sharing using a natural language processing pipeline for deidentification: Moving beyond HIPAA Safe Harbor identifiers. *JAMIA Open*, *4*(3), Article ooab069. https://doi.org/10.1093/jamiaopen/ooab069

Hesse, B. W. (2018). Can psychology walk the walk of open science? *American Psychologist*, *73*(2), 126–137. https://doi.org/10.1037/amp0000197

Inter-university Consortium for Political and Social Research. (2023). *About ICPSR*. https://www.icpsr.umich.edu/web/pages/

Joel, S., Eastwick, P. W., & Finkel, E. J. (2018). Open sharing of data on close relationships and other sensitive social psychological topics: Challenges, tools, and future directions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 86–94. https://doi.org/10.1177/2515245917744281

Kaiser, K. (2009). Protecting respondent confidentiality in qualitative research. *Qualitative Health Research*, *19*(11), 1632–1641. https://doi.org/10.1177/1049732309350879

Kapiszewski, D., & Karcher, S. (2021). Transparency in practice in qualitative research. *PS: Political Science & Politics*, *54*(2), 285–291. https://doi.org/10.1017/S1049096520000955

Karcher, S., Kirilova, D., Pagé, C., & Weber, N. (2021). How data curation enables epistemically responsible reuse of qualitative data. *The Qualitative Report*, *26*(6), 1996–2010. https://doi.org/jeslib.2021.1208

Kemmis, S., McTaggart, R., & Nixon, R. (2015). Critical theory and critical participatory action research. In H. Bradbury (Ed.), *The SAGE handbook of action research* (pp. 453–464). Sage.

Kennedy, D. M. (2012). *Don't shoot: One man, a street fellowship, and the end of violence in inner-city America*. Bloomsbury.

Kleinberg, B., Mozes, M., van der Toolen, Y., & Verschuere, B. (2017). *NETANOS - Named entity-based text anonymization for open science*. https://osf.io/w9nhb/

Klofas, J., Hipple, N. K., & McGarrell, E. (Eds.). (2010). *The new criminal justice: American communities and the changing world of crime control*. Routledge.

Kuula, A. (2011). Methodological and ethical dilemmas of archiving qualitative data. *IASSIST Quarterly*, *34*(3–4), 12–17. https://iassistquarterly.com/public/pdfs/iqvol34_35_kuula.pdf

Levenstein, M. C., & Lyle, J. A. (2018). Data: Sharing is caring. *Advances in Methods and Practices in Psychological Science*, *1*(1), 95–103. https://doi.org/10.1177/2515245918758319

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Lincoln, Y. S., Lynham, S. A., & Guba, E. G. (2018). Paradigmatic controversies, contradictions, and emerging confluences, revisited. In N. Denzin & Y. Lincoln (Eds.), *The Sage handbook of qualitative research* (5th ed., pp. 108–150). Sage.

Lipsky, L. V. (2009). *Trauma stewardship: An everyday guide to caring for self while caring for others*. Berrett-Koehler Publishers.

Logan, T. K., & Cole, J. (2011). Exploring the intersection of partner stalking and sexual abuse. *Violence Against Women*, *17*(7), 904–924. https://doi.org/10.1177/1077801211412715

MacQueen, K. M., McLellan-Lemal, E., Bartholow, K., & Milstein, B. (2008). Team-based codebook development: Structure, process, and agreement. In G. Guest & K. M. MacQueen (Eds.), *Handbook for team-based qualitative research* (pp. 119–136). AltaMira Press/Rowman & Littlefield.

McNutt, M. (2014). Journals unite for reproducibility. *Science*, *343*(6168), 229. https://doi.org/10.1126/science.1250475

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. https://doi.org/10.1177/2515245917747656

Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook* (4th ed.). Sage.

Mozersky, J., Friedrich, A. B., & DuBois, J. M. (2022). A content analysis of 100 qualitative health research articles to examine researcher-participant relationships and implications for data sharing. *International Journal of Qualitative Methods*, *21*, 1–9. https://doi.org/10.1177/16094069221105074

Mozersky, J., Parsons, M., Walsh, H., Baldwin, K., McIntosh, T., & DuBois, J. M. (2020). Research participant views regarding qualitative data sharing. *Ethics & Human Research*, *42*(2), 13–27. https://doi.org/10.1002/eahr.500044

Mozersky, J., Walsh, H., Parsons, M., McIntosh, T., Baldwin, K., & DuBois, J. M. (2020). Are we ready to share qualitative research data? Knowledge and preparedness among qualitative researchers, IRB members, and data repository curators. *IASSIST Quarterly*, *43*(4), 1–23. https://doi.org/10.29173/iq952

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, D. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . . Yarkoni, T. (2015). SCIENTIFIC STANDARDS: Promoting an open research culture. *Science*, *348*(6342), 1422–1425. https://doi.org/10.1126/science.aab2374

Parry, O., & Mauthner, N. S. (2004). Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *Sociology*, *38*(1), 139–152. https://doi.org/10.1177/0038038504039366

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Sage.

Rallis, S. F. (2015). When and how qualitative methods provide credible and actionable evidence: Reasoning with rigor, probity, and transparency. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 137–156). Sage.

Ruggiano, N., & Perry, T. E. (2019). Conducting secondary analysis of qualitative data: Should we, can we, and how? *Qualitative Social Work*, *18*(1), 81–97. https://doi.org/10.1177/14733250177007

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage.

Siegel, J. A., Calogero, R. M., Eaton, A. A., & Roberts, T. A. (2021). Identifying gaps and building bridges between feminist psychology and open science. *Psychology of Women Quarterly*, *45*(4), 407–411. https://doi.org/10.1177/03616843211044494

Stark, E. (2009). *Coercive control: The entrapment of women in personal life*. Oxford University Press.

Steltenpohl, C. N., Lustick, H., Meyer, M. S., Lee, L. E., Stegenga, S. M., Standiford Reyes, L., & Renbarger, R. L. (2023). Rethinking transparency and rigor from a qualitative open science perspective. *Journal of Trial and Error*. https://doi.org/10.36850/mr7

Tolich, M. (2004). Internal confidentiality: When confidentiality assurances fail relational informants. *Qualitative Sociology*, *27*, 101–106.

Tracy, S. J. (2019). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact* (2nd ed.). John Wiley & Sons.

Tsai, A. C., Kohrt, B. A., Matthews, L. T., Betancourt, T. S., Lee, J. K., Papachristos, A. V., Weiser, S. D., & Dworkin, S. L. (2016). Promises and pitfalls of data sharing in qualitative research. *Social Science & Medicine*, *169*, 191–198. https://doi.org/10.1016/j.socscimed.2016.08.004

VandeVusse, A., Mueller, J., & Karcher, S. (2022). Qualitative data sharing: Participant understanding, motivation, and consent. *Qualitative Health Research*, *32*(1), 182–191. https://doi.org/10.1177/10497323211054058

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

Walsh, C. G., Xia, W., Li, M., Denny, J. C., Harris, P. A., & Malin, B. A. (2018). Enabling open-science initiatives in clinical psychology and psychiatry without sacrificing patients' privacy: Current practices and future challenges. *Advances in Methods and Practices in Psychological Science*, *1*(1), 104–114. https://doi.org/10.1177/2515245917749652

Wood, L. (2015). Hoping, empowering, strengthening: Theories used in intimate partner violence advocacy. *Affilia*, *30*(3), 286–301. https://doi.org/10.1177/0886109914563157

Yardley, S. J., Watts, K. M., Pearson, J., & Richardson, J. C. (2014). Ethical issues in the reuse of qualitative data: Perspectives from literature, practice, and participants. *Qualitative Health Research*, *24*(1), 102–113. https://doi.org/10.1177/1049732313518373